2015 IEEE 18th International Conference on Computational Science and Engineering

Automatic Human Action Recognition from Video using Hidden Markov Model

Palwasha Afsar
ALGORITMI Research Center
Department of Information Systems
University of Minho
4804-533 Guimaraes, Portugal
Email: palo_afsar77@yahoo.com

Paulo Cortez
ALGORITMI Research Center
Department of Information Systems
University of Minho
4804-533 Guimaraes, Portugal
Email: pcortez@dsi.uminho.pt

Henrique Santos
ALGORITMI Research Center
Department of Information Systems
University of Minho
4804-533 Guimaraes, Portugal
Email: hsantos@dsi.uminho.pt

***Abstract*—Posture classification is a key process for evaluating the behaviors of human being. Computer vision techniques can play a vital role in automating the overall process, however, occlusions, cluttered environment and illumination changes can make the whole task difficult. Using multiple cameras and warping known object appearance into the occluded view can solve the occlusion problem. In this paper, we present an automatic human detection and action recognition system using Hidden Markov Model and bag of Words. Background subtraction is performed using Gaussian mixture model. The algorithm is able to perform robust detection in the cluttered environment and severe occlusions. The novelty of this work is the dataset used. A private dataset has been created for this research at university of Minho. The experimental results show promising results.**

***Keywords*—Video data, Human action, Hidden Markov model, Video analysis, Video databases**

I. Introduction

The area of human action recognition can be linked to many other disciplines that analyze human motion from videos. The recognition of human basic actions (e.g., walking, sitting, jumping, waving hands) from monocular images and videos is an important task in several applications such as human computer interaction, video content retrieval and surveillance. Several methods have been proposed for human action recognition. A detailed survey can be found in [1]. Recently many research works focus on recovery of human pose, which is considered an important step of view-invariant human pose recognition. However, 3D pose reconstruction from a single viewpoint is a challenging problem because of the large number of parameters and the ambiguity caused by perspective projection [2].

For some activities, there is a huge change in performance. For instance, walking movements can vary in rate and step length. Likewise, there are anthropometric contrasts between individuals. Comparable perceptions can be made for different actions, particularly for non-cyclic activities or activities that are adjusted to environment (e.g. guiding towards a certain location, avoiding obstacles while taking a walk). A decent human activity recognition methodology ought to have the capacity to sum up over varieties inside of one class and recognize activities of distinctive classes. For expanding quantities of action classes, this will be additionally difficult as the overlap between classes will be higher. In some domains, dissemination over class labels may be a suitable option. The environment in which the action execution is performed is a vital source in the recording. Individual restriction may demonstrate harder in dynamic or cluttered environments. Besides, parts of the individual may be impeded in the recording. Lighting conditions can further impact the presence of the individual.

Observing the same action from different angles can also lead to different observations. Assuming a known camera perspective limits the utilization to static cameras. At the point when various cameras are utilized, viewpoint issues and issues with impediment can be solved, particularly when perceptions from numerous perspectives can be consolidated into a steady representation. Dynamic backgrounds increase the complication of finding an individual in the image and vigorously watching the motion. At the point when utilizing a moving camera, these difficulties turn out to be considerably harder. In vision-based human action recognition, every one of these issues ought to be tended to expressly.

Hidden Markov Model has been successively used for speech recognition in training based approaches. For action recognition, HMM transforms the problem into pattern recognition. [3] was the first to use HMM for action recognition. In their work, HMM was used to recognize six different tennis strokes among three players. Some recent works [2], [4], [5] have also shown that HMM performs well for action recognition. HMM is a stochastic state transition model that models action recognition by training. For an action recognition, the HMM which best matches an action is chosen. It achieves high recognition rates and requires low processing time.

In this paper, we propose an action recognition method based on HMM using the Bag of Words method. Time-sequential images of human action are transformed into feature vectors. All of these feature vectors are stored in a codebook where each symbol corresponds to an action by using Vector Quantization (VQ). For the training phase, the model parameters of HMM are tuned well for description of an action to achieve high results. The Model, which matches to a symbol, is selected as a recognized entity. The framework of the approach can be seen in Figure 1.

The paper is organized as follow: section 2 contains related work on human action recognition. In section 3, a brief introduction of Hidden Markov Model is given. Section 4 covers the dataset created for this research. The methodology

978-1-4673-8297-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSE.2015.41

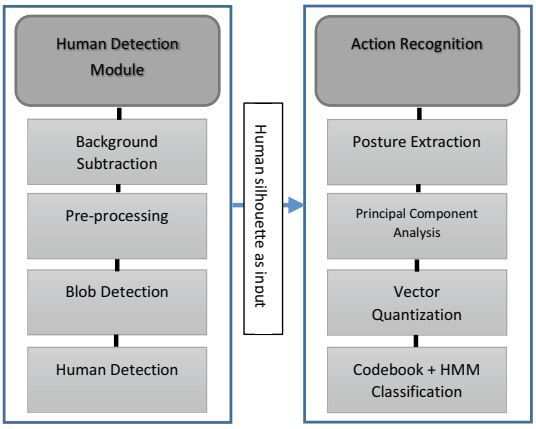

Fig. 1: Pictorial representation of the system

of the system is defined in section 5. Experimental results are discussed in section 6 and finally section 7 outlines the conclusion and future work.

II. Related Work

Researchers have explored different methods for action recognition. In 1975, Johansson's experiment shows that human can recognize activity with extremely compact observers [6]. However, in recent studies [7] proposed to use star skeleton for extracted from silhouettes for motion analysis. They used Hidden Markov based methodology for learning actions. Star skeleton is a fast skeletonization technique achieved by connecting centroid of the target object to contour extremes. For using star skeleton as a feature, the feature vector is defined as a five dimensional vector. An action is composed of a series of star skeletons and hence is are transformed into feature vector sequences. A posture codebook is designed where a feature vector is transformed into a symbol sequence. This symbol sequence is used by HMM for recognizing an action. The algorithm achieves robust recognition rates.

[8] Used joint position and pose angles for recognition of simple action like walking, standing and waving hands. The high dimensional 3D data joint space data is decomposed into a feature space where each feature corresponds to motion of single or multiple joints. Hidden Markov model (HMM) is used for learning of actins together with Adaboost classifier. The results obtained on 20 actions shows the effectiveness of the approach.

[4] work focus on posture classification based on projection histogram with HMM for assuring temporal coherence of the posture. For occlusion handling, the single camera postures are passed to multi-camera system to give robust classification. In contrast to [4] a stationary camera was used by [9] for recognizing 15 different continuous human activities in real-time. The activities recognized includes waving hands, raising hands, sitting down and bendind down. Activities were described as a continuous sequence of discrete postures, derived from affine invariant descriptor. SVM is use for classification of results.

[5] Developed a video surveillance system capable of recognizing human posture from video. The system consists of two modules: human detection and posture classification. In human detection module, human blob are extracted from the background using an adaptive background module base don mixture of Gaussians. For the variation in human pose, pseudo 2D hidden Markov models (P2HMM) is used for representing and recognizing human postures based on its 2D elastic matching property. For classification, observation sequences are extracted from current image and human blobs are classified as the human posture with the highest likelihood.

[10] proposed a system for recognizing human daily life activities. The method utilizes a hierarchical structure of actions and describes it as a tree. The actins are modeled using Hidden Markov model, which output action as a time series feature vectors. The recognition process starts at the root and moves on to the recognition of the child nodes. Hierarchical recognition offers several advantages like simplification of low-level models, recognition of various levels of abstraction, and response to novel data by decreasing the degree of data details. Results show that their proposed algorithm can recognize some of the actions.

[3] A feature-based bottom-up approach is proposed by yamato. HMM was applied to one set of time-sequential images that were transformed into feature vector. These features were converted into symbols using Vector Quantization. The parameters of HMM were optimized well for training actions categories. The HMM which best matches an action category was chosen. They achieve recognition rates higher than 90% on real-time sequential images of sport sequences.

The posture of human is considered important for recognition of human actions. Inspired from [7], we propose an action recognition algorithm using HMM representation of postures extracted from human silhouettes. Posture contour is normalized to achieve a standard size. The observation vectors are extracted from these silhouettes and Vector Quantization (VQ) is used to map features into symbols. A posture codebook representing actions is build and each symbol correspond to a different action stored in the codebook. The existing research works used actions captured against a uniform background with little or no occlusion. Also the tests set contains single subject performing an action. Our method achieve robust results on real life actions in the presence of severe occlusion and bad weather condition like wind, raining.

III. Hidden Markov Model (HMM)

Hidden Markov Model (HMM) is a statistical Markov Model in which the system being model is assumed to be a Markov process with hidden states as shown in Figure 2. An HMM can also be modeled as dynamic Bayesian network. Each state of the HMM outputs a state transition symbol. While we can observe the output symbol of HMM, we cannot observe the states of HMM. The formal definition of HMM is as follows:

$$\lambda = \{A, B, \pi\}$$

X is the state alphabet set

$$X = \{x_1, x_2, x_3..x_n\}$$

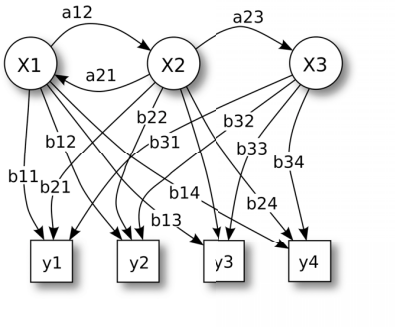

Fig. 2: Diagramtic representation of HMM with three states

and Y is the observation set

$$Y = \{y_1, y_2, y_3, ..y_n\}$$

We define Q to be a fixed state sequence of length T, and corresponding observation O:

$$Q = \{q_1, q_2, q_3..q_t\} \quad O = \{o_1, o_2, o_3...o_t\}$$

A is a transition array, storing the probability of state j following state i. Note that the state transition probabilities are independent of time:

$$A = [a_ij], a_ij = P(q_t = x_j | q_{t-1} = x_i).$$

B is the observation array, storing the probability of observation k being produced from the state j, independent of time t:

$$B = [b_i(k)], b_i(k) = P(x_t = y_k | q_t = s_i)$$

π is the initial probability array.

$$\pi = [\pi_i], \pi_i = P(q_1 = x_i)$$

Two assumptions are made by the model. The first, called the Markov assumption, states that the current state is dependent only on the previous state, this represents the memory of the model:

$$P(q_t | q_1^{t-1}) = P(q_t | q_{t-1})$$

The independence assumption states that the output observation at time t is dependent only on the current state, it is independent of previous observations and states:

$$P(o_t | O_1^{t-1}, q_1^t) = P(o_t | q_t)$$

HMM performs actioin recognition in two parts: training the model and computing the probability that the observation sequence was generrated by the model λ . Each model is trained so that it is able to generate the symbols for training data. The training process can be further optimized using parameters (A,B,π).

Fig. 3: Installation of cameras for capturing the dataset at UMinho

Fig. 4: Some images from the dataset

IV. Dataset

For this research, a video dataset has been created at university of Minho. The dataset was recorded from 0800h to 1900h for a total of 7 days. Therefore, varying light conditions can be observed in the dataset. It also includes videos having rainy and windy weather conditions that make it quite a challenging dataset. For capturing the dataset, two cameras HIK Vision and IR Network were installed, as shown in Figure 3. The environment for capturing the video dataset is not controlled. The cameras are installed behind the window glass, therefore, a reflection is always present. The dataset is comprised of hundreds of small videos (with a few seconds to a few minutes each) that correspond to 32GB. The type of actions recorded includes walking, running, group interactions, talking on mobile, drinking coffee, standing and shaking hands Figure 4. All of these actions are real life moments. The angle of the camera is set in such a way that it always has two or three trees framing the scene. This has been done to make the dataset challenging and test the efficiency of the proposed algorithm.

V. Methodology

In current video surveillance system, majority of work is devoted to background modeling and object tracking. However, the recognition of human activities has not gained enough attention. The work in this paper focuses on two modules i.e., human detection in presence of occlusion and classification of activities e.g., walking, sitting. The adaptive background model is used to detect moving objects from video streams. The background model is based on mixture of Gaussians and is robust to illumination and clutter.

VI. Overview of the system

The detection of moving object is an important component for many computer applications e.g., including action recognition, surveillance systems etc. The system proposed in this paper is based on two main modules i.e., Human detection Module and action recognition module. The human detection module uses a background subtraction algorithm based on Gaussian mixture model. The algorithm compares a color or grayscale video frame to a background model to determine whether individual pixels are part of the background or the moving object. A foreground mask is then computed. Morphological operators are applied to the resulting foreground mask to eliminate any noise present. Blob analysis is performed to look for connected regions or pixels that most likely corresponds to moving objects.

For action recognition, all subsequent frames representing an action are collected. Human boundary is considered one of the best ways to represent an action. Posture contour is extracted from the silhouettes of human and normalization is performed to achieve a standard size. As the boundary of human consists of many neighboring pints that are almost the same, Principal Component analysis is performed for dimensionality reduction. The observation vectors are extracted from these silhouettes and Vector Quantization (VQ) is used to map features into symbols. A posture codebook representing actions is build and each symbol correspond to a different action stored in the codebook. An extracted feature vector is mapped to a codebook symbol and the output is a hence a sequence of posture symbols. Hidden Markov model is used for training different actions that optimize model parameters and recognition is performed by choosing maximum likelihood.

A. Background Segmentation

The main concept behind background subtraction is to subtract the image from a reference image that models the background scene. The main steps of the algorithm are as follow:

1) Background modeling: constructs a reference image representing the background.

2) Threshold selection: determines appropriate threshold value to be used in subtraction operation to achieve a desired detection rate.

3) Subtraction operation: It is also called pixel classification and is used for classification type of a given pixel i.e., the pixel is part of the background or the moving object.

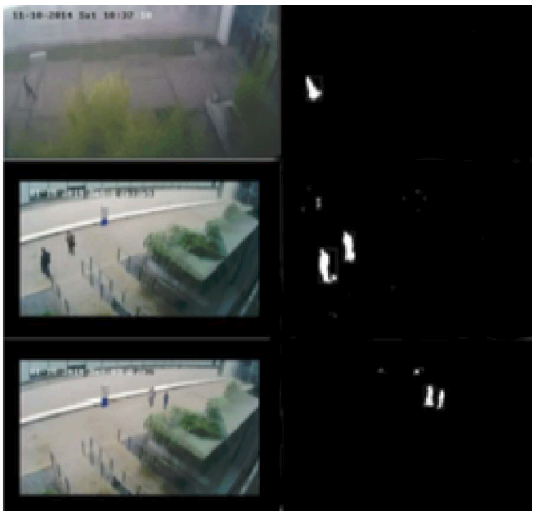

Fig. 5: Detection results from Human detection module

B. Blob detection

In blob detection, all of the foreground pixels in are grouped into disconnected blobs. A blob can represent a) no object b) part of moving object c) a single moving object with possible foreground trail and d) multiple moving objects. The first one refers to the presence of ghosts that are caused by shadows. In this work our aim is to remove the entire ghosts to achieve better classification results. The second one is due to aperture problem that are produced when an object is starting to move. These partial blobs are maintained because eventually they will grow to full blobs. Majority of blobs fall into third category. The last case of multiple objects occurs when people start walking in a group and are close to one another. These large blobs finally split into multiple small blobs when a group is dispersed.

VII. Experimental Results

To evaluate the performance of our approach, the algorithm was tested on two types of actions i.e., walking and sitting. For these action classes, history of both of the actions were recorded over time and all frames that can represent either of the actions were selected. The testing was performed on single frames representing an action and a series of frames. The videos were recorded by HIK Vision and IR Network cameras at 1080x1920 HD pixel resolution. The size of the image was reduced to 576x720 resolution and preprocessing was performed to remove any kind of noise present. A set of data that was not included in training was used for testing. The algorithm was able to detect human in cluttered environment as shown in Figure 5. For recognition of actions, the test frames were compared with the one, stored in codebook. The recognition results can be seen in Table I and Table I.

VIII. Conclusion

We have presented an algorithm for automatically detecting human from videos and recognizing actions in presence of occlusions and illumination changes. Human posture is a

TABLE I: Confusin matrix on dataset1 with average accuracy of 97%

Known	Sit	Walk
Sit	1.00	0.00
Walk	0.06	0.94

TABLE II: Confusin matrix on dataset2 with average accuracy of 95%

Known	Sit	Walk
Sit	0.99	0.01
Walk	0.09	0.91

key feature for recognizing any kind of actions. The paper focuses on using the boundary of human as a main feature of recognizing actions. For obtaining the foreground image, background subtraction was performed using Gaussian of mixture model. For each of the action class, a number of frames representing an action were recorded. Feature vectors are extracted from silhouettes and Vector Quantization (VQ) is used to map features into symbols. Action recognition is done using Hidden Markov Model (HMM). Our algorithm shows promising results on two datasets with average accuracy of 95% and 97%. However, there are some limitations of the current work. The system was tested only on two datasets. In our future work, we aim to check the robustness of the proposed work on more actions. We will be also improving the algorithm to work on joint information of a human skeleton for action recognition.

ACKNOWLEDGMENT

This work is funded by the Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia) within the Project Scope UID/CEC/00319/2013 and research grant SFRH/BD/84939/2012.

REFERENCES

[1] P. Afsar, P. Cortez, and H. Santos, "Automatic visual detection of human behavior: a review from 2000 to 2014," *Expert Systems with Applications*, 2015.

[2] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[3] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*. IEEE, 1992, pp. 379–385.

[4] R. Cucchiara, A. Prati, and R. Vezzani, "Posture classification in a multi-camera indoor environment," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 1. IEEE, 2005, pp. I–725.

[5] L. H. W. Aloysius, G. Dong, H. Zhiyong, and T. Tan, "Human posture recognition in video sequence using pseudo 2-d hidden markov models," in *Control, Automation, Robotics and Vision Conference, 2004. ICARCV 2004 8th*, vol. 1. IEEE, 2004, pp. 712–716.

[6] G. Johansson, "Visual motion perception." *Scientific American*, 1975.

[7] H.-S. Chen, H.-T. Chen, Y.-W. Chen, and S.-Y. Lee, "Human action recognition using star skeleton," in *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*. ACM, 2006, pp. 171–178.

[8] F. Lv and R. Nevatia, "Recognition and segmentation of 3-d human action using hmm and multi-class adaboost," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 359–372.

[9] V. Kellokumpu, M. Pietikäinen, and J. Heikkilä, "Human activity recognition using sequences of postures." in *MVA*, 2005, pp. 570–573.

[10] T. Mori, Y. Segawa, M. Shimosaka, and T. Sato, "Hierarchical recognition of daily human actions based on continuous hidden markov models," in *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*. IEEE, 2004, pp. 779–784.